

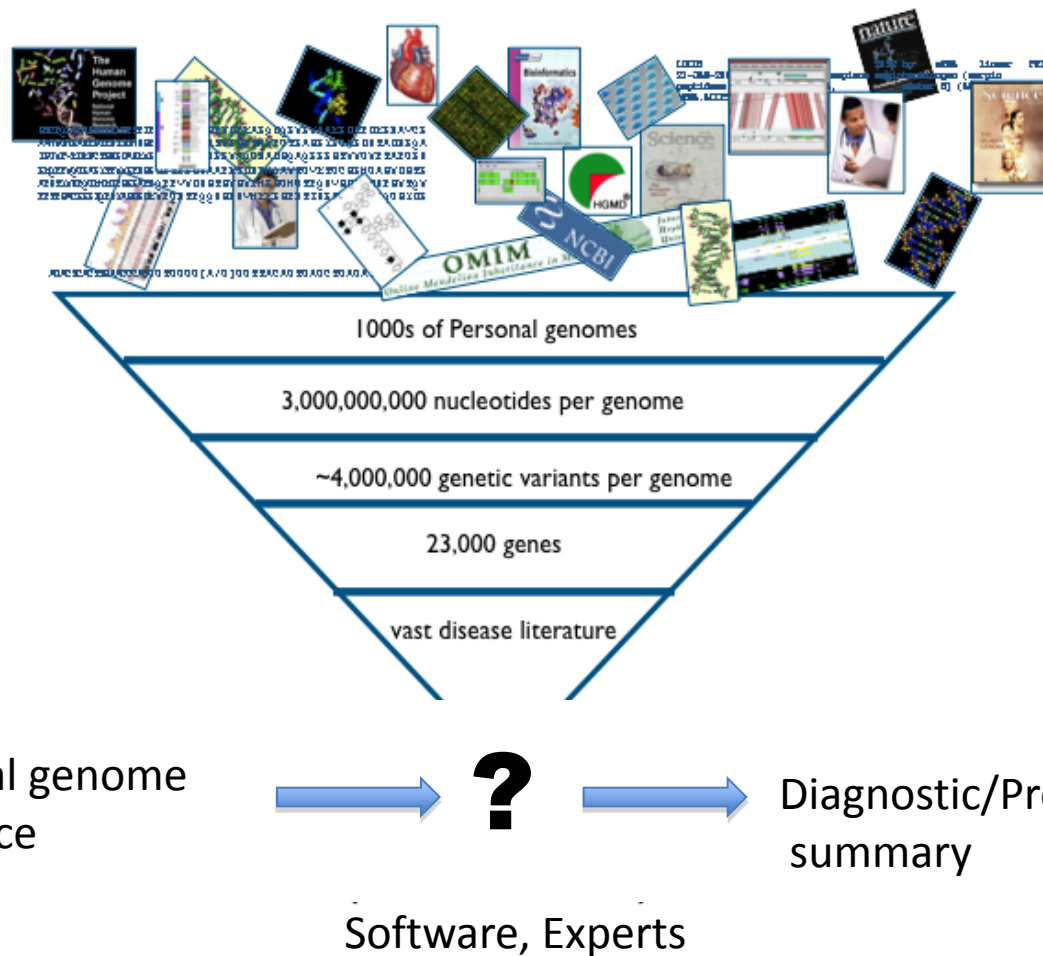
PREPARING AND MANAGING WHOLE GENOME SEQUENCE DATA FOR THE CLINICAL SETTING

Martin G. Reese, PhD

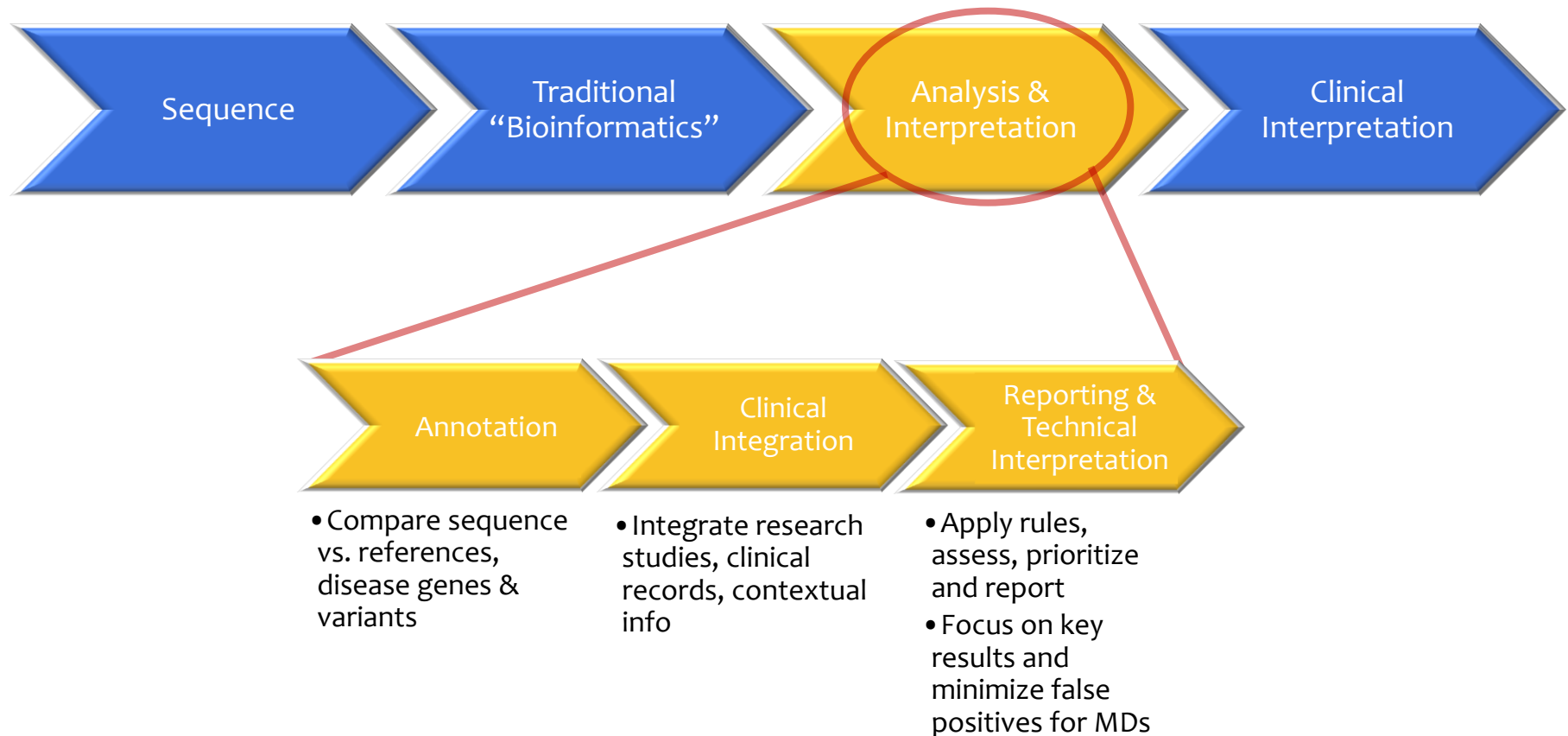
SACGHS, Washington, DC

June 15th, 2010

The Challenge: How to distill a genome's worth of sequence into clinically actionable information?



Workflow of WGS in the Clinic



Topics

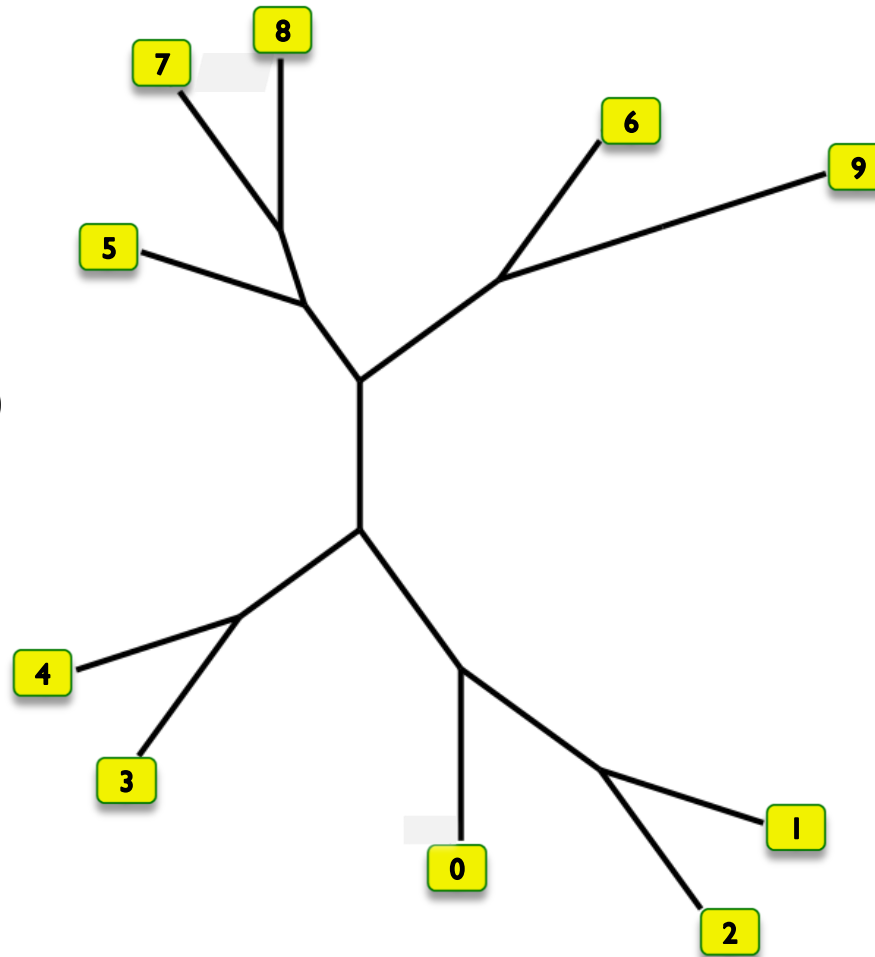
- Genome quality assessment and control
- Integrated system's approach
- Clinical interpretation

The “10Gen” Data Set

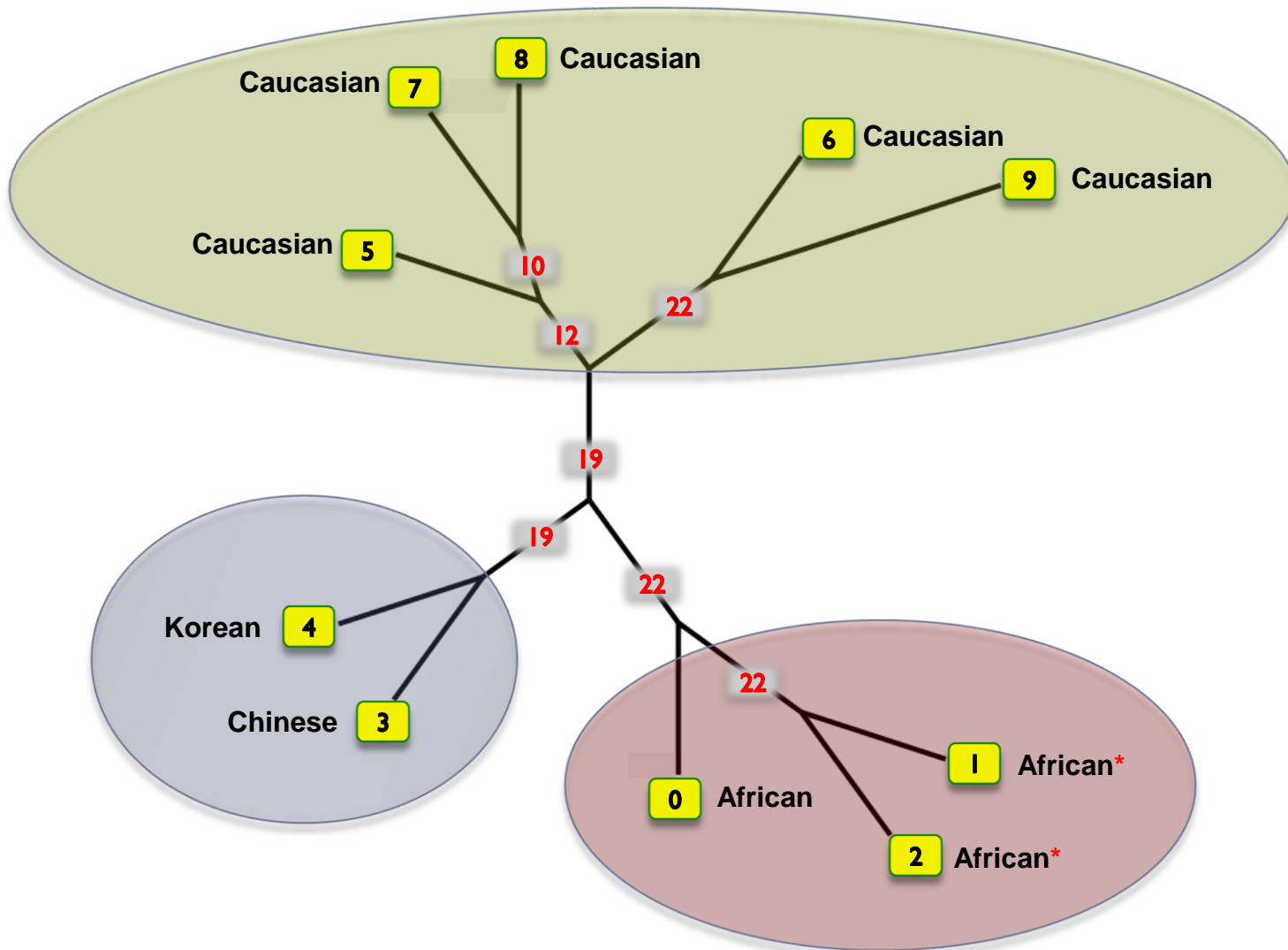
Genome	Individual	Ethnicity	Platform	Reference
0	NA19240	African	Life SOLiD	De la Vega, et al. 2009
1	NA18507	African	Illumina	Bentley et al. 2008
2	NA18507	African	Life SOLiD	McKernan, et al. 2009
3				
4	Chinese	Asian	Illumina	Wang et al. 2008
5	Korean	Asian	Illumina	Ahn et al. 2009
6	Venter	Caucasian	Sanger	Levy et al. 2007
7	Watson	Caucasian	Roche 454	Wheeler et al. 2007
8	NA07022	Caucasian	CGenomics	Drmanac, et al. 2009
9	NA12878	Caucasian	Life SOLiD	De la Vega, et al. 2009
	Quake	Caucasian	Helicos	Pushkarev et al. 2009

The Data Are Structured

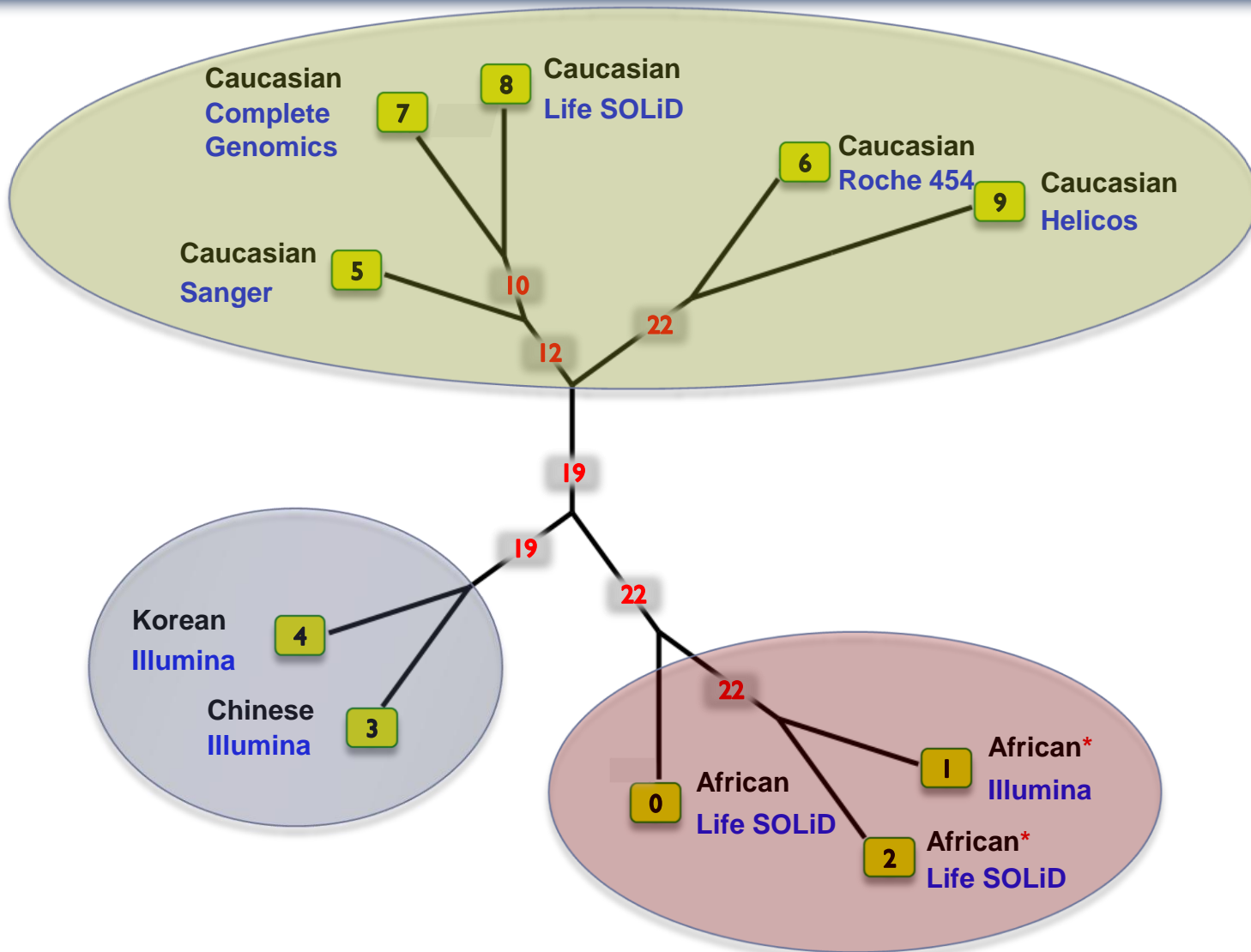
$$d = \frac{N_s - (N_s \cap N_L)}{N_s}$$



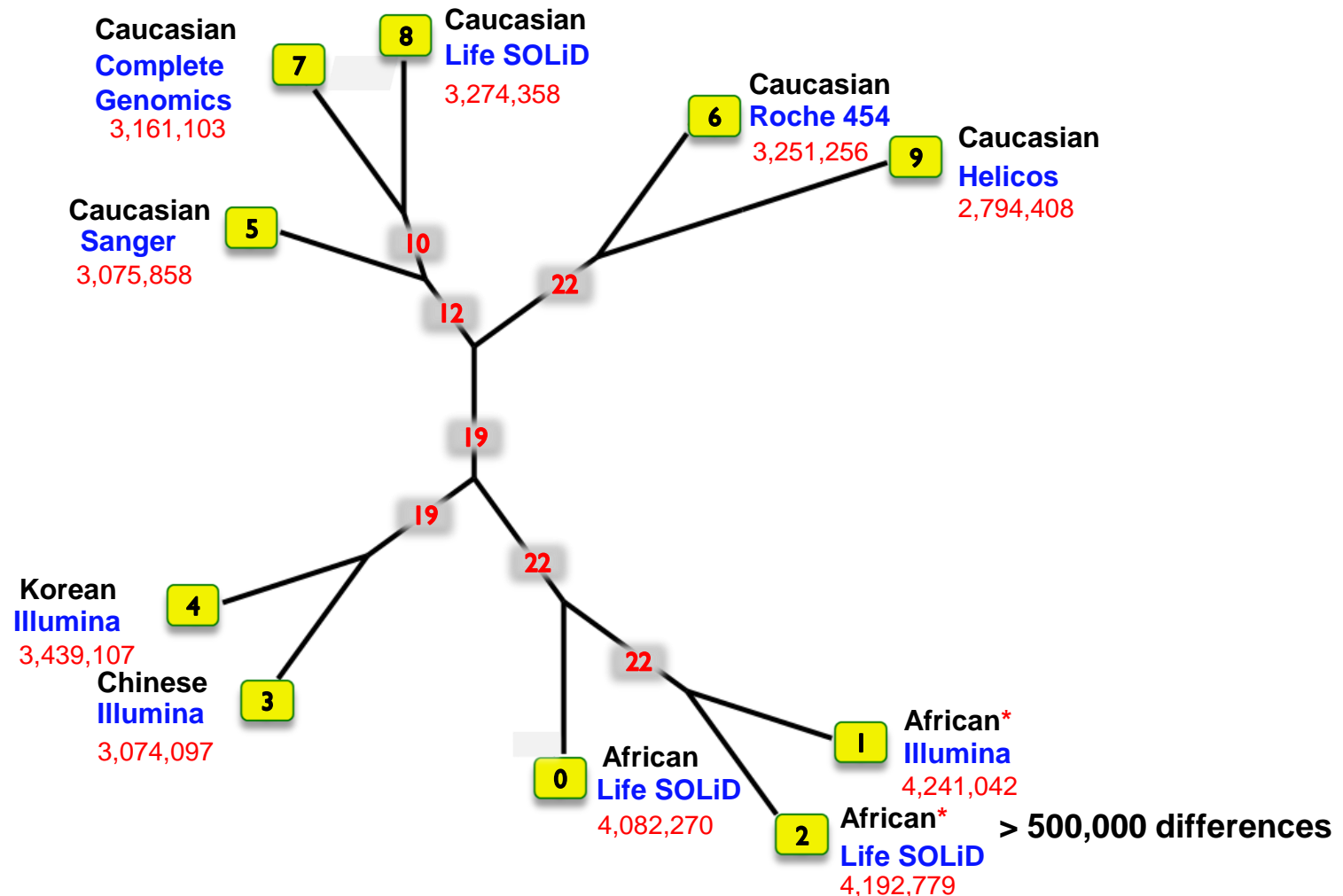
There Is a Strong Historical Signal



Technologies Give Consistent Results



Over 500,000 single nucleotide variants (SNVs) differences within NA18507 genome



The Challenge Of Personal Genome Analysis

- ▶ Average variant file contains 3.5+ million variants incl. SNVs, indels, structural variants (SVs)
- ▶ ~21,000 coding variants
- ▶ How do we make sense of data at this scale?
 - ▶ Need new tools to tackle this problem

Using Established Resources



Genes with allelic variant data	2,099
Allelic variants	15,772
Amino acid variants	10,361
Intron variants	1,086

Each Genome Contains About 100 OMIM alleles

Number of OMIM alleles/genome

Genome	Heterozygous	Homozygous
0	51	30 (+16)
1	57	33 (+17)
2	56	33 (+18)
3	55	30 (+14)
4	68	24 (+14)
5	61	21 (+17)
6	53	16 (+20)
7	72	22 (+18)
8	67	20 (+13)
9	78	17 (+13)

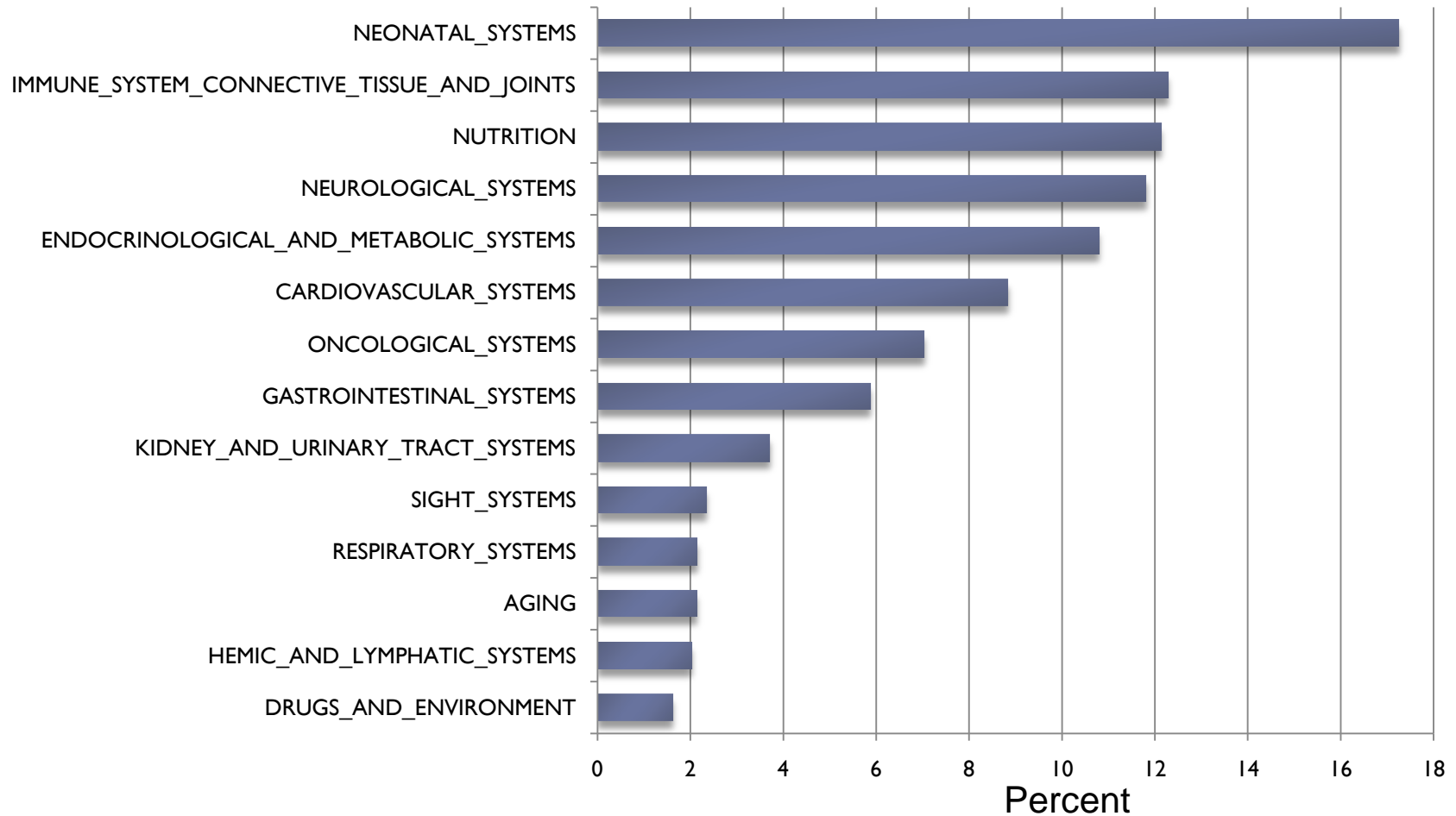
Homozygous Alleles: No Childhood Fatal Disorder Variants

Genome	Diseases (selected)
0	Protection against coronary heart disease (CHD), macular degeneration
1	Susceptibility to inflammatory bowel disease (IBD), macular degeneration, drug interaction
2	Susceptibility to IBD, macular degeneration
3	Congestive heart failure, Preeclampsia
4	T2Diabetes, hypertension, congestive heart failure, macular degeneration, Preeclampsia, CHD
5	Diabetes mellitus type I, congestive heart failure, drug interactions, leptin polymorphisms
6	T2Diabetes, tuberculosis, drug interactions, color perception
7	Lumbar disk disease, hypercholesterolemia, CHD
8	Congestive heart failure, hypercholesterolemia
9	Crohn's disease, CHD

OMIM Not Sufficient For Whole Genomes

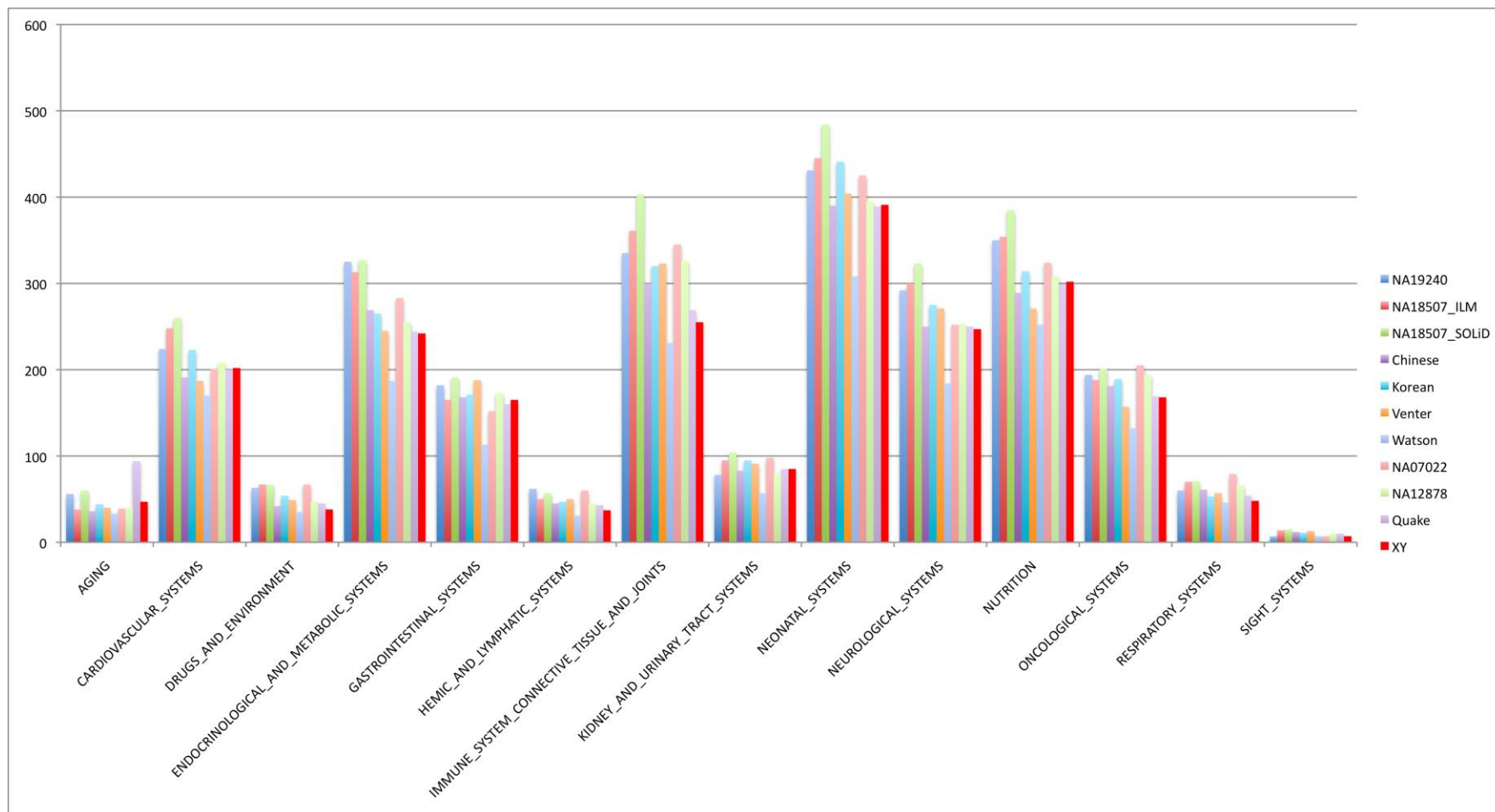
- ▶ Average genome contains:
 - ▶ ~ 21,000 coding SNVs
 - ▶ ~10,000 protein sequence changing (non-synonymous) SNVs
 - ▶ Only ~100 of these are known OMIM alleles
- ▶ How do we analyze the remaining 99%?
 - ▶ Need new tools to tackle this problem
 - ▶ Software pipeline for automatic variant annotation.

3,626 Genes Classified (Omicia Disease Gene Database v2.3)

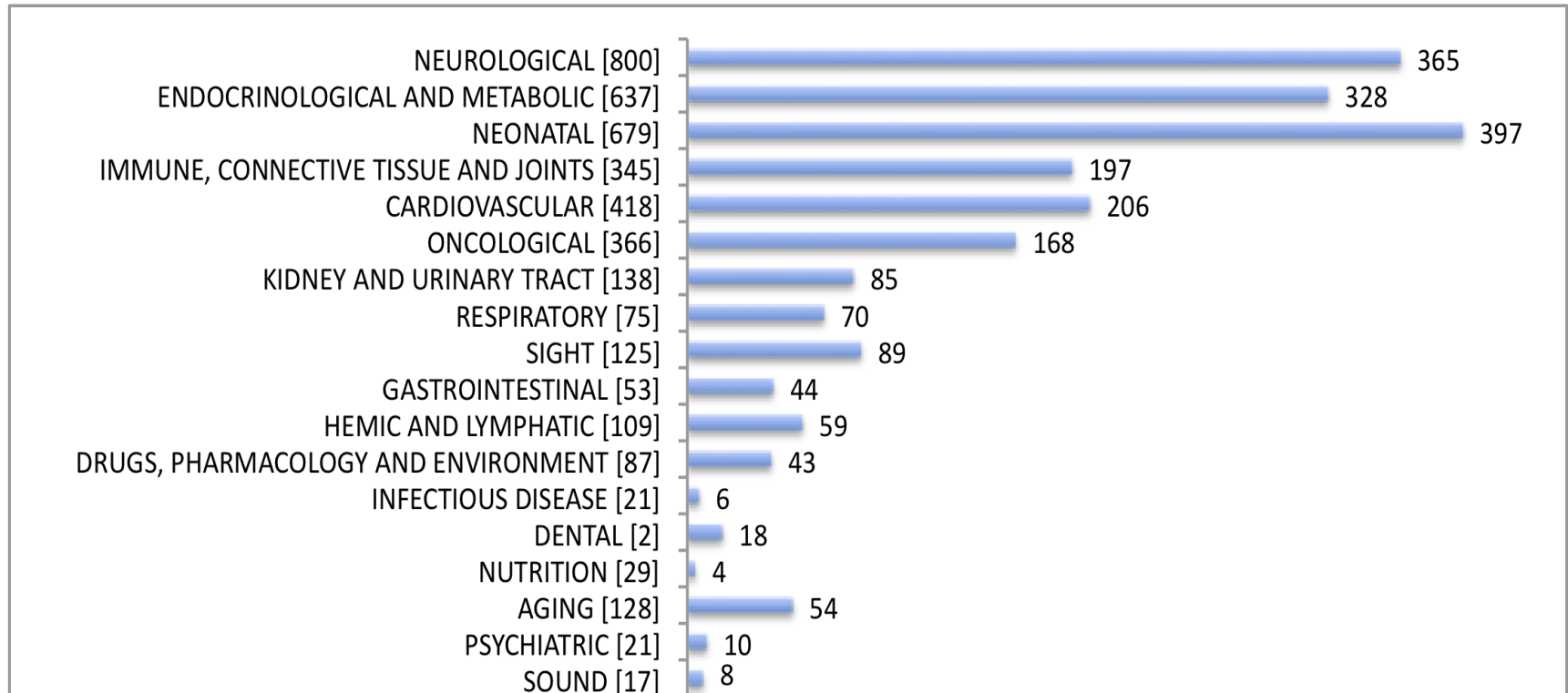


¹ Harrison's "Principles of Internal Medicine", 12th Edition

Distribution of Non-Synonymous Variants in 10Gen Dataset



Distribution of 2,179 Protein Sequence Changing Variants in Disease Genes for Typical Genome



Criteria for Ranking and Interpretation

- Quality of sequenced variant
- Zygosity
- Protein function assignment
 - Loss-of-function mutation (stop codon, frameshift indel)
 - Non-synonymous mutation
 - Splice site mutation
- Overlap with documented pathogenic variant in literature
 - OMIM (15,000+ variants)
 - HGMD (~75,000+ variants)
 - GWAS (From NHGRI website; updated weekly)
 - PharmGKB (3,425 pharmacogenomics variants as of June 6, 2010)

Areas of Value for the Individual and Physician



- Disease Predisposition
 - Examples: Cardiovascular Disease, Cancer, Diabetes, Obesity
- Adverse Drug Reactions
 - Examples: Anesthesia, Antidepressants, Asthma Medication
- Drug responders
 - Examples: Metabolism, Cancer
- Prenatal Risks
 - Understanding genetic risks for your children
 - Parents may be carriers of genetic diseases such as Tay-Sachs, Cystic Fibrosis

Three Very Recent Examples of WGS in the Clinic

nature

Vol 456 | 6 November 2008 | doi:10.1038/nature07485





ARTICLES

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley^{1,2,3,4,*}, Elaine R. Mardis^{2,3,*}, Li Ding^{2,3}, Bob Fulton³, Michael D. McLellan³, Ken Chen³, David Dooling³, Brian H. Dunford-Shore³, Sean McGrath³, Matthew Hickenbotham³, Lisa Cook³, Rachel Abbott³, David E. Larson³, Dan C. Koboldt³, Craig Pohl³, Scott Smith³, Amy Hawkins³, Scott Abbott³, Devin Locke³, LaDeana W. Hillier^{3,8}, Tracie Miner³, Lucinda Fulton³, Vincent Magrini^{2,3}, Todd Wylie³, Jarret Glasscock³, Joshua Conyers³, Nathan Sander³, Xiaoli Shi³, John R. Osborne³, Patrick Minx³, David Gordon⁸, Asif Chinwalla³, Yu Zhao¹, Rhonda E. Ries¹, Jacqueline E. Payton⁵, Peter Westervelt^{1,4}, Michael H. Tomasson^{1,4}, Mark Watson^{3,4,5}, Jack Baty⁶, Jennifer Ivanovich^{4,7}, Sharon Heath^{1,4}, William D. Shannon^{1,4}, Rakesh Nagarajan^{4,5}, Matthew J. Walter^{1,4}, Daniel C. Link^{1,4}, Timothy A. Graubert^{1,4}, John F. DiPersio^{1,4} & Richard K. Wilson^{2,3,4}

THE LANCET

Clinical assessment incorporating a personal genome

Dr Euan A. Ashley MRCP , Atul J. Butte MD , Matthew T. Wheeler MD , Rong Chen PhD , Teri F. Klein PhD , Frederick F. Dewey MD , Joel T. Dudley , Kelly E. Ormond MSc , Aleksandra Pavlovic BS , Alexander A. Morgan MS , Dmitry Pushkarev , Norma F. Neff PhD , Prof. Louanne Hudgins MD , Li Gong PhD , Laura M. Hodges PhD , Dorit S. Berlin PhD , Caroline F. Thorn PhD , Katrin Sangkuhl PhD , Joan M. Hebert MA , Mark Woon BSE , Hersh Sagreliya , Ryan Whaley BS , Joshua W. Knowles MD , Michael F. Chou PhD , Joseph Y. Thakuria MD , Abraham M. Rosenbaum PhD , Alexander Wait Zarenek PhD , George M. Church PhD , Prof. Henry T. Greely JD , Stephen R. Quake PhD , Prof. Russ B. Altman MD

Summary

Background

The cost of genomic information has fallen steeply, but the clinical translation of genetic risk estimates remains unclear. We aimed to undertake an integrated analysis of a complete human genome in a clinical context.

Methods

We assessed a patient with a family history of vascular disease and early sudden death. Clinical assessment included analysis of this patient's full genome sequence, risk prediction for coronary artery disease, screening for causes of sudden cardiac death, and genetic counselling. Genetic analysis included the development of novel methods for the integration of whole genome and clinical risk. Disease and risk analysis focused on prediction of genetic risk of variants associated with mendelian disease, recognised drug responses, and pathogenicity for novel variants. We queried disease-specific mutation databases and pharmacogenomics databases to identify genes and mutations with known associations with disease and drug response. We estimated post-test probabilities of disease by applying likelihood ratios derived from integration of multiple common variants to age-appropriate and sex-appropriate pre-test probabilities. We also accounted for gene-environment interactions and conditionally dependent risks.

Findings

Analysis of 2.6 million single nucleotide polymorphisms and 752 copy number variations showed increased genetic risk for myocardial infarction, type 2 diabetes, and some cancers. We discovered rare variants in three genes that are clinically

Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing

Jared C. Roach^{1,*}, Gustavo Glusman^{1,*}, Arian F. A. Smit^{1,*}, Chad D. Huff^{1,2,*}, Robert Hubley¹, Paul T. Shannon¹, Lee Rowen¹, Krishna P. Pant³, Nathan Goodman¹, Michael Bamshad⁴, Jay Shendure⁵, Radoje Drmanac³, Lynn B. Jorde², Leroy Hood^{1,†} & David J. Galas^{1,†}

We analyzed the whole-genome sequences of a family of four, consisting of two siblings and their parents. Family-based sequencing allowed us to delineate recombination sites precisely, identify 70% of the sequencing errors (resulting in >99.999% accuracy), and identify very rare single-nucleotide polymorphisms. We also directly estimated a human intergenerational mutation rate of $\sim 1.1 \times 10^{-8}$ per position per haploid genome. Both offspring in this family have two recessive disorders: Miller syndrome, for which the gene was concurrently identified, and primary ciliary dyskinesia, for which causative genes have been previously identified. Family-based genome analysis enabled us to narrow the candidate genes for both of these Mendelian disorders to only four. Our results demonstrate the value of complete genome sequencing in families.

Whole-genome sequences from four members of a family represent a qualitatively different type of genetic data than whole-genome sequences from individuals

or sets of unrelated genomes. They enable inheritance analyses that detect errors and permit the identification of precise locations of recombination events. This leads in turn to near-complete knowledge of inheritance states and haplotypes power analyses that include the identification of genomic features with nonclassical inheritance patterns, such as hemizygous deletions or copy number variants (CNVs). Identification of inheritance patterns in the pedigree permits the detection of ~70% of sequencing errors and sharply reduces the search space for

disease-causing variants. These analyses would be far less powerful in studies that had fewer markers (such as standard genotype or exome data sets) or that had sequences from fewer family members.

DNA from each family member was extracted from peripheral blood cells and sequenced at CGI (Mountain View, California) with a nanoarray-based short-read sequencing-by-ligation technology (1), including an adaptation of the pairwise end-sequencing strategy (2). Reads were mapped to the National Center for Biotechnology Information (NCBI) reference genome (fig. S1 and tables S1 and S2). Polymorphic markers used for this analysis were single-nucleotide polymorphisms (SNPs) with at least two variants among the four genotypes of the family, averaging 802 base pairs (bp) between markers. We observed 4,471,510 positions at which at least one family member had an allele that varied from the reference genome. This corresponds to a Watterson's theta (θ_w) of 9.5×10^{-4} per site for the two parents and the reference sequence (3), given the fraction of the genome successfully genotyped in each parent (fig. S1). This is a close match to the estimate of $\theta_w = 9.3 \times 10^{-4}$ that we obtained by combining two previously published European genomes and the reference sequence (4). Of the 4.5 million variant positions, 3,665,772 were variable within the family; the rest were homozygous and identical in all four members. Comparisons to known SNPs show that 323,255 of these 3.7 million SNPs are novel.

For each meiosis in a pedigree, each base position in a resulting gamete will have inherited one of two parental alleles. The number of inheritance patterns of the segregation of alleles in

¹Institute for Systems Biology, Seattle, WA 98103, USA.

²Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84109, USA.

³Complete Genomics, Inc. (CGI), Mountain View, CA 94043, USA.

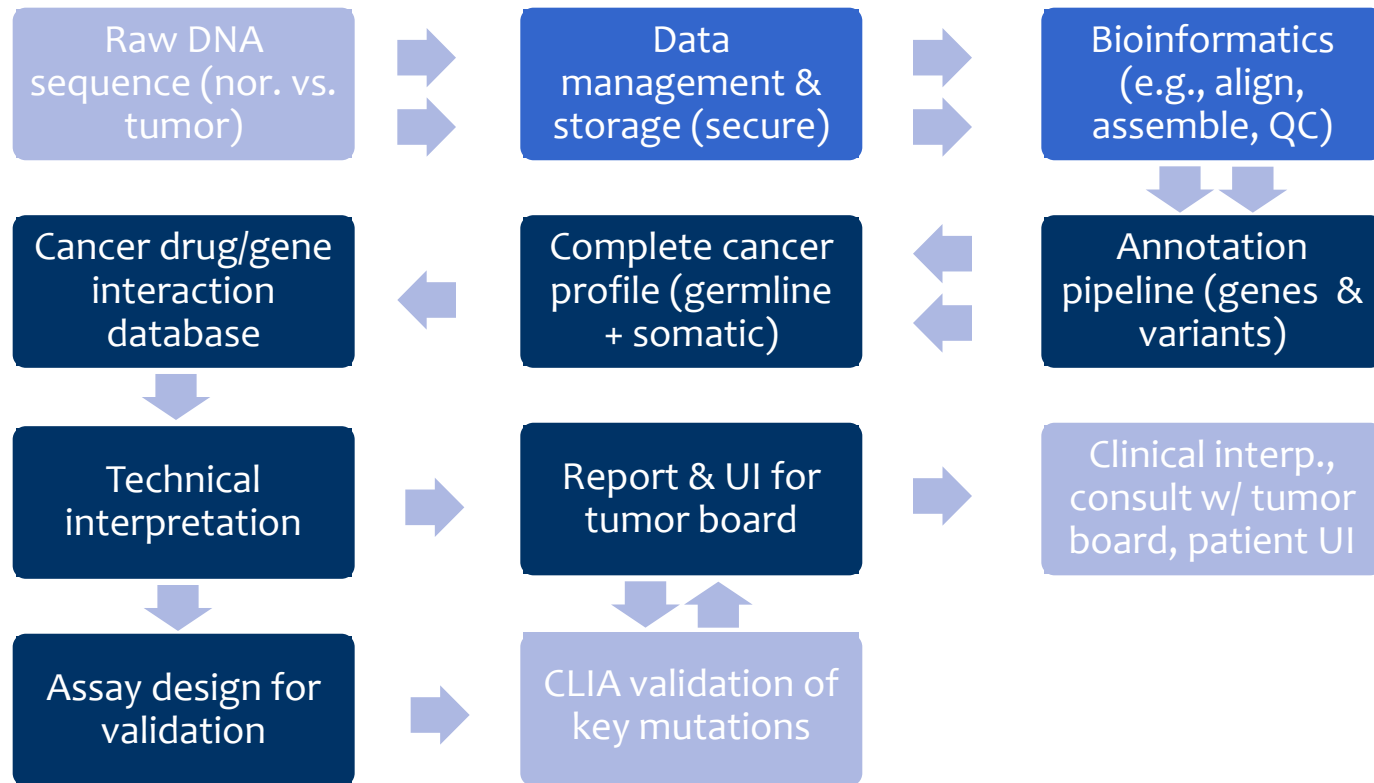
⁴Department of Pediatrics, University of Washington, Seattle, WA 98195, USA.

⁵Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA.

*These authors contributed equally to this work.

[†]To whom correspondence should be addressed. E-mail: dgalas@systemsbiology.org (D.J.G.); lhood@systemsbiology.org (L.H.)

Example: Genomic Cancer Care Alliance: Proposed Interpretation Workflow



Example: Pharmacogenomics: Warfarin/VKCOR1

Gene	Variant	rsID	risk concern	reference	Low dose required		High dose	XY genotype (=high dose)	# reads
VKCOR1	1173G>A	rs9934438	Low dose group at risk for severe bleeding	G	A	A	G	G/G	22
	-1639C>T	rs9923231		C	T	T	C	C/C	19
	1542C>G	rs8050894		C	G	G	C	C/C	13
	-497A>C	rs2884737		A	A	C	A	A/A	20
	2255G>A	rs2359612		G	A	A	G	G/G	9

Conclusion:

PG1002 genotypes for known function altering variants are for the high dose Warfarin allele, and therefore no risk.

Take home messages (1)

- Quality assessment and control
 - Identity testing (ie genotyping validation experiment)
 - Sequence quality differs in genomic regions (error rate reported on DNA chip validation experiments (99.98% concordance) is underestimate; specific issues with indels, structural rearrangements, etc.)
 - Verification experiments (orthogonal technology validation in CLIA lab)
 - Data security
- Integrated systems for technical interpretation are needed that link literature information with medical practice
 - Variant ranking by quality and phenotype severity
 - Rule-based decision support systems needed (transparency critical)
 - Re-interpretation necessary if new discoveries are made

Take home messages (2)

- Clinical interpretation
 - Layered reporting critical
 - Technical genetic reports -> Genomic medicine expert -> Treating physician
 - Interpretation within genetic/environmental background
 - Focused interpretation based on clinical parameters and family history
 - Integration into EMRs

Acknowledgements

Yandell Lab



Barry Moore

Carson Holt

Hao Hu

Deepak Anthony

Mark Yandell

Marth Lab



Gabor Marth

Francisco de la Vega

Kevin McKernan



Fidel Salas

Edward S. Kiruluta

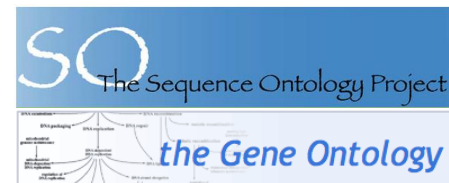
Archie Russell

George Miklos

Paul Billings

Erwin Frise

Martin Reese



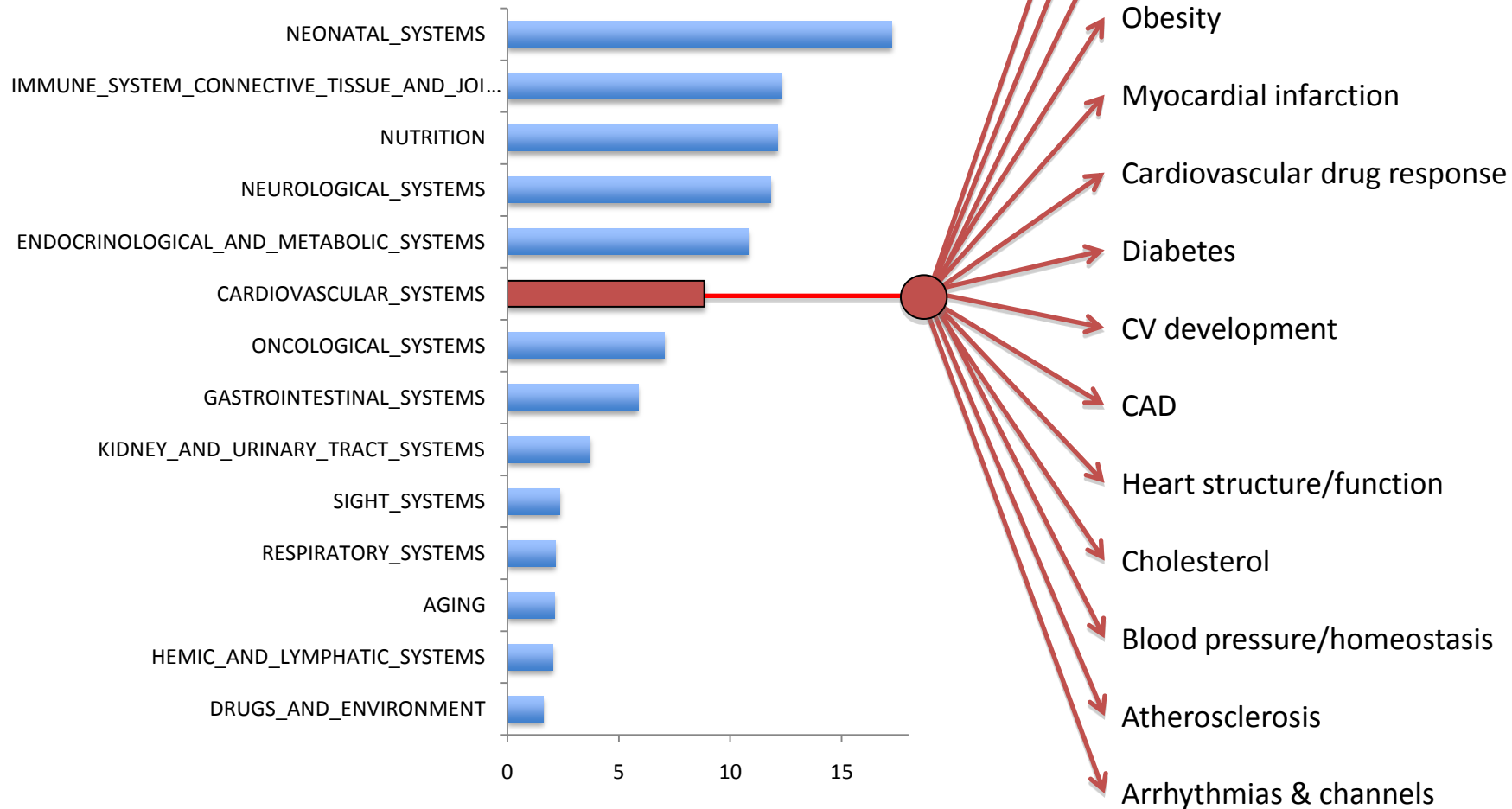
Karen Eilbeck

Chris Mungall

This work was supported by NIH SBIR grants **1R44HG003667** to Omicia/Yandell, SBIR **1R43HG002993** and **1R44HG002991** to Omicia and an NIH ARRA GO grant **1RC2HG005619** to Yandell/Omicia all administered by the National Human Genome Research Institute (**NHGRI**).

Back-up slides

Sub-Classifications of CVD genes within the Omicia Disease Gene Ontology version 3.0.



Omicia Disease Gene Database (version 2.4)

- Rolling up disease gene ontologies into Harrison derived categories
 - *Harrison Internal Medicine*
- 3,738 curated human disease genes from most prominent sources that include
 - Omicia hand-selected from research publications
 - OMIM (Online Mendelian Inheritance in Man)
 - HGMD (Human Genome Mutation Database)
 - GWAS studies (NHGRI web server)
 - Published Human Disease Gene set by Jimenez-Sanchez (2001) (913 genes)
- Each gene published in the literature as playing a causative role in one or more human diseases
 - Two levels of evidence(i.e. family-genetic study, functional assays, biochemical assay, significant GWAS, etc)
- Annotated with clinical phenotype information

Relationship Among Genomic Sequence Pioneers

